

Bayesian Probabilities

Frank Dellaert

January 2002

Copyright (c) 2002 Frank Dellaert

1



The concepts in this lecture
are **deceptively** simple !

In fact, most people never really understand
them, even after years of practice...

**However, if you really grasp them,
they will be immensely valuable !**

January 2002

Copyright (c) 2002 Frank Dellaert

2

Outline

Knowledge as Probability

Conditional Probabilities

Probability Densities

Likelihood & Bayes Law

Bayes Classifier

Naïve Bayes Classifier

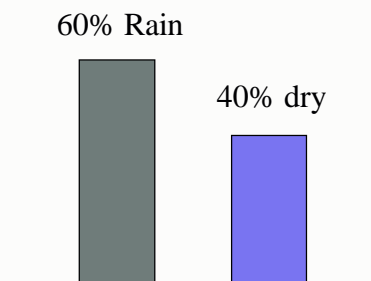
January 2002

Copyright (c) 2002 Frank Dellaert

3

The Bayesian Paradigm

- Knowledge as a probability distribution



January 2002

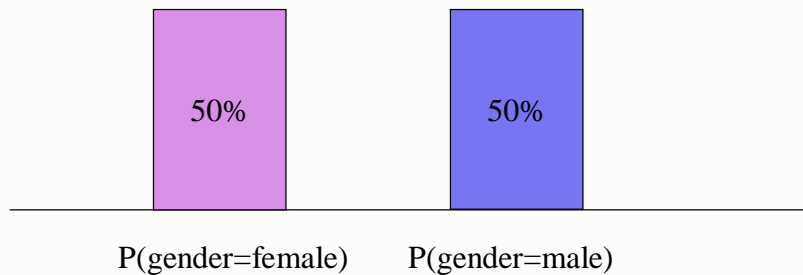
Copyright (c) 2002 Frank Dellaert

4

Probability of an Event

What is our KNOWLEDGE about...

Discrete attribute: gender



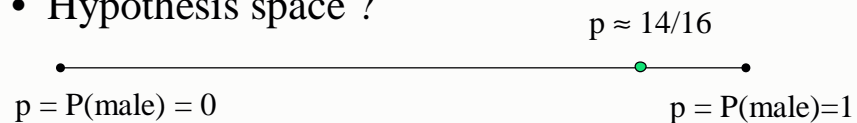
January 2002

Copyright (c) 2002 Frank Dellaert

5

The Simplest Learner Ever !

- No Input Attributes
- Discrete Target Concept: **gender in class**
- Hypothesis space ?



- Learning algorithm = ?
- $p \approx \text{\#positive examples} / \text{\#examples}$

January 2002

Copyright (c) 2002 Frank Dellaert

6

Output of simple learner ?

- Model = (unconditional) $p = P(\text{male}) \approx 14/16$
- Learner output ?

$$\operatorname{argmax}_{\text{class}}(p(\text{class})) = \text{male} !$$

Pretty lame ! But it will get better !

Training and Test

- Training Set:

- Test Set:

- Sample error ?
- True error ?

Outline

Knowledge as Probability

Conditional Probabilities

Probability Densities

Likelihood & Bayes Law

Bayes Classifier

Naïve Bayes Classifier

January 2002

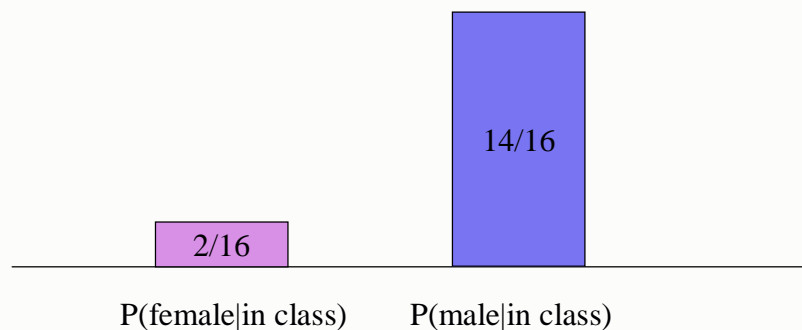
Copyright (c) 2002 Frank Dellaert

9

Conditional Probability

Knowledge about attributes given the class, e.g.

“How probable is attribute gender **given that he/she is in class** ?”



January 2002

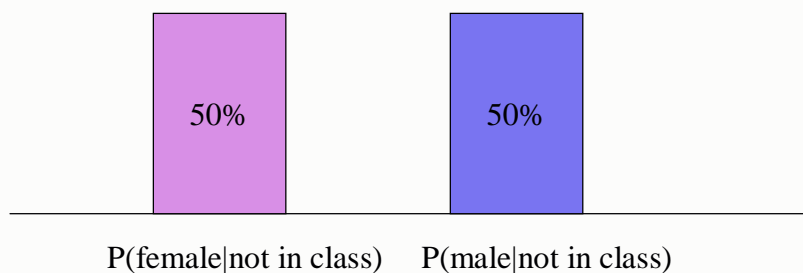
Copyright (c) 2002 Frank Dellaert

10

Conditional Probability

“How probable is gender given that he/she is **not in class** ?”

Discrete attribute: gender

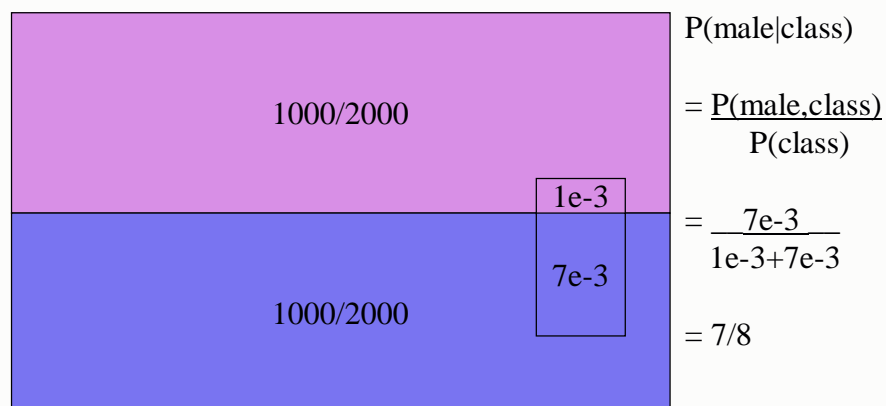


January 2002

Copyright (c) 2002 Frank Dellaert

11

Conditional Probability Picture



January 2002

Copyright (c) 2002 Frank Dellaert

12

Conditional Probability

Conditional probability $P(\text{gender}|\text{class})$ is a function of ?



What is $P(\text{male}|\text{class}) + P(\text{female}|\text{class})$?
100% !!

What is $P(\text{male}|\text{class}) + P(\text{male}|\text{not in class})$?
NOT A PROBABILITY !!

January 2002

Copyright (c) 2002 Frank Dellaert

13

Learning Conditional Probabilities

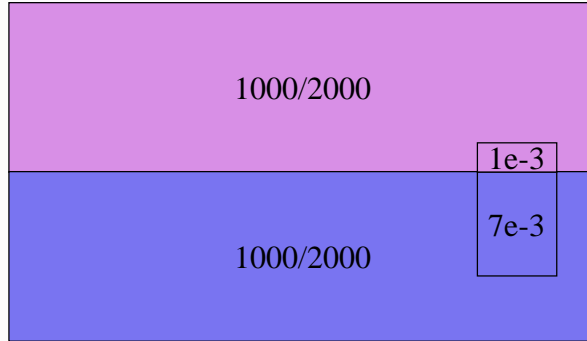
- Input Attribute: in class/not in class
- Discrete Target Concept: **gender**
- Hypothesis space ?
 $(p_0, p_1) \in \mathbb{R}^2$
- Learning algorithm = ?
- $p_0 \approx \frac{\#\text{positive}|\text{ not in class}}{\#\text{examples}|\text{ not in class}}$
- $p_1 \approx \frac{\#\text{positive}|\text{ in class}}{\#\text{examples}|\text{ in class}}$

January 2002

Copyright (c) 2002 Frank Dellaert

14

Joint Probability Distribution



Counts:			
	not in class	in class	
female	998	2	1000
male	986	14	1000
	1984	16	



Probabilities			
	not in class	in class	
female	49.90%	0.10%	50.00%
male	49.30%	0.70%	50.00%
	99.20%	0.80%	

January 2002

Copyright (c) 2002 Frank Dellaert

15

Conditional Distributions

Probabilities			
	not in class	in class	
female	49.90%	0.10%	50.00%
male	49.30%	0.70%	50.00%
	99.20%	0.80%	

Target: gender
Input: class

Divide !

Target: class
Input: gender

P(gender class)			
	not in class	in class	
female	50.30%	12.50%	62.80%
male	49.70%	87.50%	137.20%
	100.00%	100.00%	

P(class gender)			
	not in class	in class	
female	99.80%	0.20%	100.00%
male	98.60%	1.40%	100.00%
	198.40%	1.60%	

January 2002

Copyright (c) 2002 Frank Dellaert

16

Brute Force MAP Hypothesis

- Model =

P(gender class)		
	not in class	in class
male	49.70%	87.50%

- Learner output ?

$$\text{target(inputs)} = \underset{\text{class}}{\text{argmax}} P(\text{target|inputs})$$

- gender(not in class) ?
- gender(in class): ?

Posterior Probability

MAP Hypothesis

January 2002

Copyright (c) 2002 Frank Dellaert

17

The Omnipotent Joint PDF

Can answer any question, e.g.

$P(\text{Play}|\text{Sunny})$?
 $P(\sim\text{Play}|\text{Sunny},\text{Hot})$

but also
 $P(\text{Hot},\text{Humid})$
 $P(\text{Hot},\text{Humid}|\text{Play})$

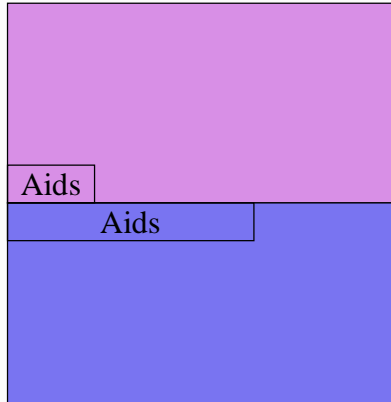
Bias ?
 None !
 Need lots of data !

Joint PDF for PlayTennis						
	Wind	Outlook	Temp	Humidity	Play	P(W,O,T,H,P)
Weak		Sunny	Hot	High	Yes	0.30%
Weak		Sunny	Hot	High	No	0.50%
Weak		Sunny	Hot	Normal	Yes	0.30%
Weak		Sunny	Hot	Normal	No	0.10%
Weak		Sunny	Mild	High	Yes	0.10%
Weak		Sunny	Mild	High	No	0.60%
Weak		Sunny	Mild	Normal	Yes	0.20%
Weak		Sunny	Mild	Normal	No	0.30%
Weak		Sunny	Cool	High	Yes	0.27%
Weak		Sunny	Cool	High	No	0.26%
Weak		Sunny	Cool	Normal	Yes	0.25%
Weak		Sunny	Cool	Normal	No	0.25%
Weak		Sunny	Hot	High	Yes	0.24%
Weak		Sunny	Hot	High	No	0.23%
Weak		Sunny	Hot	Normal	Yes	0.23%
Weak		Sunny	Hot	Normal	No	0.22%
Weak		Overcast	Mild	High	Yes	0.21%
Weak		Overcast	Mild	High	No	0.20%
Weak		Overcast	Mild	Normal	Yes	0.20%
Weak		Overcast	Mild	Normal	No	0.19%
Weak		Overcast	Cool	High	Yes	0.18%
Weak		Overcast	Cool	High	No	0.17%
Weak		Overcast	Cool	Normal	Yes	0.17%
Weak		Overcast	Cool	Normal	No	0.16%
Weak		Overcast	Hot	High	Yes	0.15%
...

January 2002

18

Medical Example



P(gender)		P(g)
gender		
male		500
female		500
		1000

P(gender,aids)			P(g,a)
gender	aids		
male	yes		20
male	no		480
female	yes		5
female	no		495
			1000

Marginal P(aids)			P(aids)
aids			
yes		25	2.50%
no		975	97.50%
		1000	100.00%

January 2002

Copyright (c) 2002 Frank Dellaert

19

Conditional Marginalization

P(gender,aids,test)			
gender	aids	test	
male	yes	pos	17
male	yes	neg	3
male	no	pos	90
male	no	neg	390
female	yes	pos	5
female	yes	neg	0
female	no	pos	95
female	no	neg	400
			1000

Conditional P(test aids)				
aids	test	P(aids,test)	P(aids)	P(test aids)
yes	pos	22	25	88%
yes	neg	3	25	12%
no	pos	185	975	19%
no	neg	790	975	81%
		1000		

Hypothesis aids(test)	
test	aids
pos	NO
neg	NO

Conditional P(aids test)				
aids	test	P(aids,test)	P(test)	P(aids test)
yes	pos	22	207	10.6%
no	pos	185	207	89.4%
yes	neg	3	793	0.4%
no	neg	790	793	99.6%
		1000		

We need a more accurate test !!!

January 2002

Copyright (c) 2002 Frank Dellaert

20

Frequency Format:

- 25 out of every people 1,000 have AIDS. Of these 25 people with AIDS, 22 will have a positive test. Of the remaining 975 people without AIDS, 185 will still have a positive test. Imagine a sample of people (with no symptoms) who have positive tests in your AIDS screening. How many of these people do actually have AIDS ?
- 22 out of $(22+185) = 10.6\%$
- "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats" by Gigerenzer and Hoffrage (1995), *Psychological Review*, 102, 684-704.

January 2002

Copyright (c) 2002 Frank Dellaert

21

Outline

Knowledge as Probability

Conditional Probabilities

Probability Densities

Likelihood & Bayes Law

Bayes Classifier

Naïve Bayes Classifier

January 2002

Copyright (c) 2002 Frank Dellaert

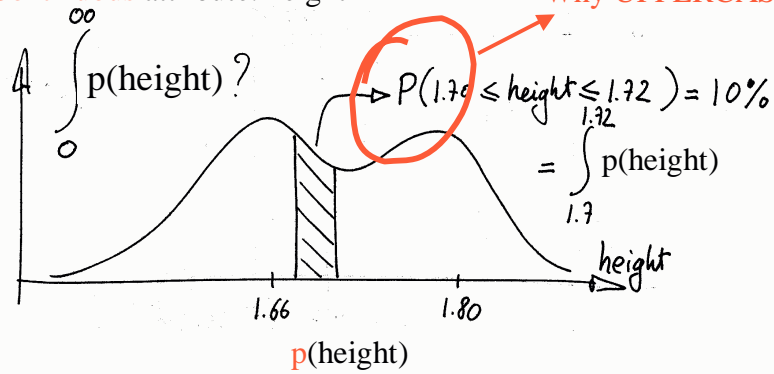
22

Probability Density

What is our KNOWLEDGE about...

Continuous attribute: height

Why UPPERCASE ?



January 2002

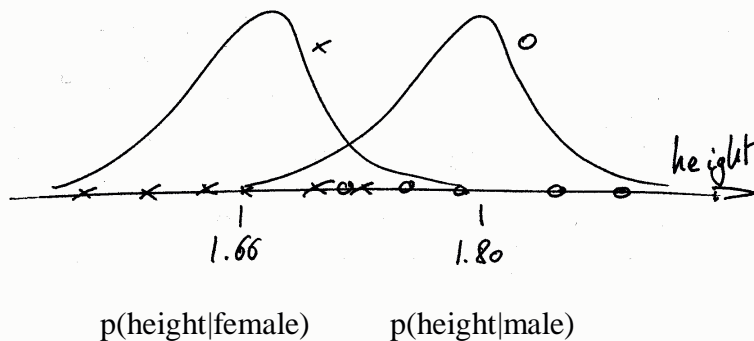
Copyright (c) 2002 Frank Dellaert

23

Conditional Density

Knowledge about features given the class, e.g.

"How probable is a certain height given that a person is female?"



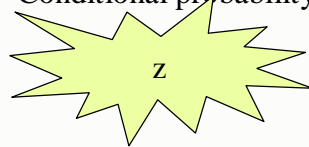
January 2002

Copyright (c) 2002 Frank Dellaert

24

Conditional Density

Conditional probability $p(z|\text{class})$ is a function of ?



class

What is $\int p(z|\text{class}) dz$?

100% !!

What is $p(z|\text{class}) + p(z|\text{not in class})$?
NOT A DENSITY !!

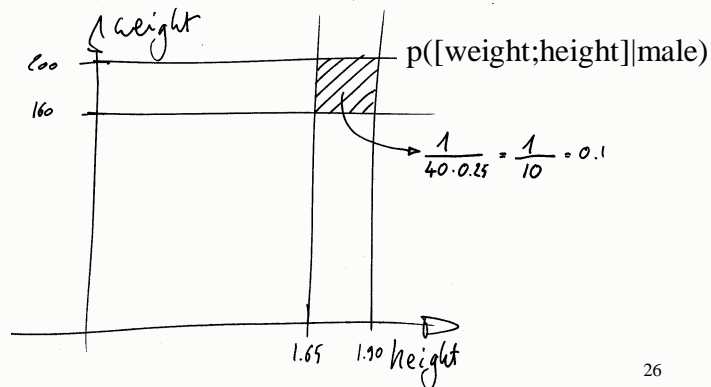
January 2002

Copyright (c) 2002 Frank Dellaert

25

Exercise

- Suppose: male adults have heights between 1.65 and 1.90, and weights between 160 and 200 lbs, **uniformly distributed**.



January 2002

26

Probability and Energy

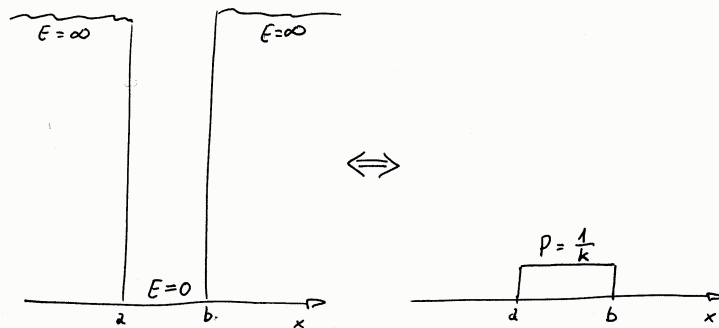
Wake up !! You are about to learn a very useful point of view

January 2002

Copyright (c) 2002 Frank Dellaert

27

Probability $\sim \exp(-\text{Energy})$



January 2002

Copyright (c) 2002 Frank Dellaert

28

Probability $\sim \exp(-\text{Energy})$

- $E > 0$
- E can be infinite
- $P(x) = 1/Z \exp(-E(x))$
where $Z = \sum_x \exp(-E(x))$
- Z = normalizing factor, makes P(x) a true probability
- Fancy name: **partition function**

January 2002

Copyright (c) 2002 Frank Dellaert

29

Origin: Statistical Physics

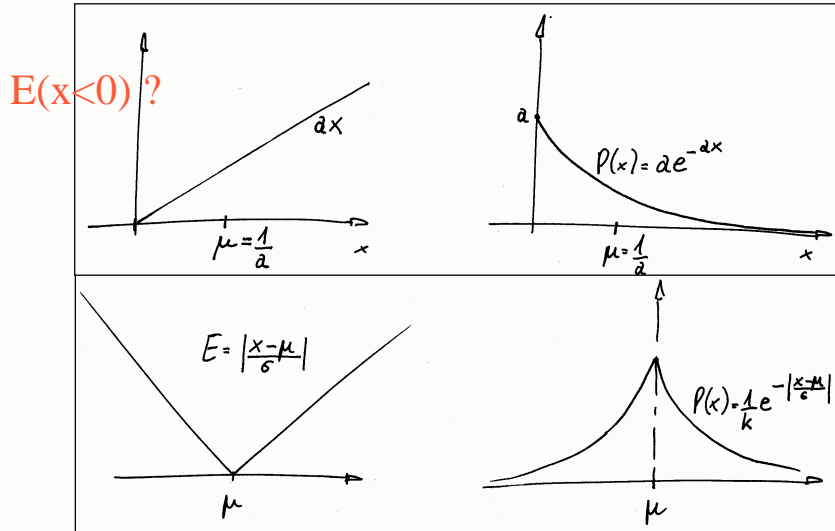
- The higher the free energy $H(x)$ of a system in state x , the less probable it is.
- The Boltzmann-Gibbs Distribution:
$$P(x) = 1/Z \exp(-H(x)/kT)$$
where $k = 1.38E-16$, $T = \text{temp. in Kelvin}$
- Z = partition function = hard to compute !

January 2002

Copyright (c) 2002 Frank Dellaert

30

Linear Energy

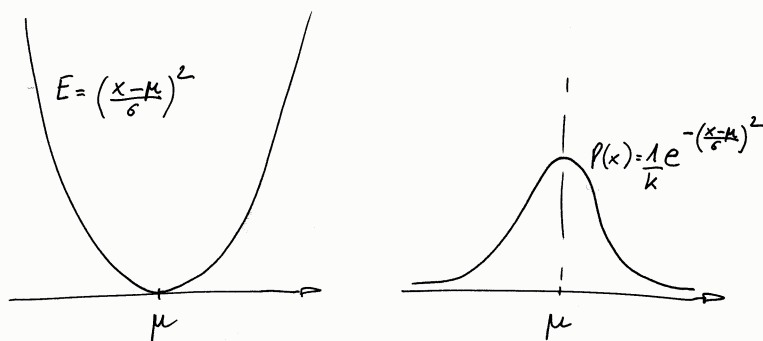


January 2002

Copyright (c) 2002 Frank Dellaert

31

A Familiar Density Exposed



The normal density (Gaussian) is the probability density function associated with the square error !

January 2002

Copyright (c) 2002 Frank Dellaert

32

Outline

Knowledge as Probability

Conditional Probabilities

Probability Densities

Likelihood & Bayes Law

Bayes Classifier

Naïve Bayes Classifier

January 2002

Copyright (c) 2002 Frank Dellaert

33

Likelihood

- Recap: $p(\text{height}|\text{gender})$: conditional density of height given gender = $f(\text{height})$
- Now: **assume height is given** (i.e. measured)
- What is most **likely** gender ?

January 2002

Copyright (c) 2002 Frank Dellaert

34

Likelihood

- Likelihood of class == Probability of measurements assuming class is correct
- More formally: if x is target class, and z is a vector of measurements, then **the likelihood of x given z** , written $L(x;z)$, is any function proportional to $P(z|x)$

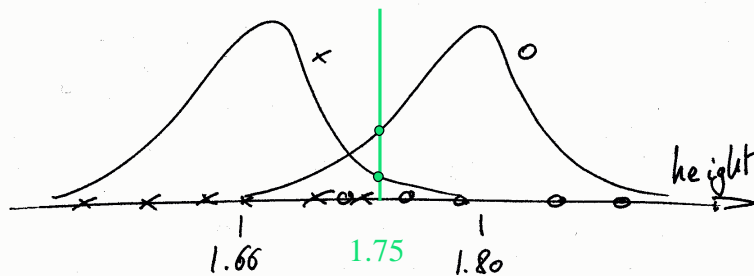
January 2002

Copyright (c) 2002 Frank Dellaert

35

Likelihood Example

How **likely** is a certain class given that a person has **this height** ?



$$L(x; \text{height}=1.75) \sim P(1.75|x)$$

$$L(o; \text{height}=1.75) \sim P(1.75|o)$$

January 2002

Copyright (c) 2002 Frank Dellaert

36

These are the single most misunderstood facts by computer scientists attempting to apply statistics to [whatever]

Given z , Likelihood $L(\text{class};z) \sim P(z|\text{class})$ is a function of ?

z

class

$L(\text{male};\text{height}=1.75) + L(\text{female};\text{height}=1.75) = ??$
 NOT 100% ! Likelihood is NOT a probability distribution
 And hence is not true knowledge (in the Bayesian sense)

January 2002

Copyright (c) 2002 Frank Dellaert

37

Discrete Example

Remember:

		Conditional $P(\text{test} \text{aids})$		
aids	test	$P(\text{aids},\text{test})$	$P(\text{aids})$	$P(\text{test} \text{aids})$
yes	pos	22	25	88%
yes	neg	3	25	12%
no	pos	185	975	19%
no	neg	790	975	81%
		1000		

Now: given that test is positive, what is...

- Likelihood of aids=yes ?

$$P(\text{test}=\text{pos}|\text{aids}=\text{yes}) = 88\%$$

- Likelihood of aids=no ?

$$P(\text{test}=\text{pos}|\text{aids}=\text{no}) = 19\%$$

January 2002

Copyright (c) 2002 Frank Dellaert

38

Maximum Likelihood

- In general: **maximum likelihood hypothesis:**
- $h_{ML} = \operatorname{argmax}_x P(\text{data}|x) = \operatorname{argmax}_x L(x;\text{data})$
- Aids example:
 - $\text{aids}_{ML}(\text{test}=\text{pos}) = \text{yes}$
 - $\text{aids}_{ML}(\text{test}=\text{neg}) = \text{no}$

January 2002

Copyright (c) 2002 Frank Dellaert

39

What's the problem ??

Remember:

Conditional P(aids test)				
aids	test	P(aids,tes)	P(test)	P(aids test)
yes	pos	22	207	10.6%
no	pos	185	207	89.4%
yes	neg	3	793	0.4%
no	neg	790	793	99.6%
		1000		

- MAP (?) hypothesis $\operatorname{argmax}(\text{aids}|\text{test})$ disagrees !
- Why ?
- So, which one is correct ??
- Answer: Decision Theory !

January 2002

Copyright (c) 2002 Frank Dellaert

40

Decision Theory

- Specify Risk(class,decision)
- Calculate $P(\text{class}|\text{data})$
- Calculate expected risk of a decision:
 - $E[\text{Risk}|\text{decision}] = \sum P(\text{class}|\text{data}) \text{Risk}(\text{class},\text{decision})$
- Best decision = $\text{argmin}_{\text{decision}} E[\text{Risk}|\text{decision}]$

Risk		decision		P(aids pos)
		yes	no	
aids	yes	0	1	10.63%
aids	no	1	0	89.37%
E[Risk]		0.89	0.11	

Risk		decision		P(aids pos)
		yes	no	
aids	yes	0	10	10.63%
aids	no	1	0	89.37%
E[Risk]		0.89	1.06	

January 2002

Copyright (c) 2002 Frank Dellaert

41

Posterior vs Likelihood

- Posterior from Joint:
 - $P(x|z) = P(x,z)/P(z)$
- Likelihood model
 - $P(z|x) = P(x,z)/P(x)$
- Chain Rule:
 - $P(z|x) = P(x,z)/P(x) \Rightarrow P(x,z) = P(z|x)P(x)$
 - $P(a,b,c) = P(a|b,c)P(b|c)P(c)$

January 2002

Copyright (c) 2002 Frank Dellaert

42

Bayes Law

- Likelihood model
 - Often easier to estimate
 - Sometimes easier to learn
- Calculate Posterior from Likelihood:
 - $P(x|z) = P(x,z)/P(z) = P(z|x) P(x) / P(z)$
- Since $P(z)$ is constant when z given:

$$P(x|z) \sim L(x;z)P(x)$$

January 2002

Copyright (c) 2002 Frank Dellaert

43

In other words...

- Knowledge **before** measurement: $P(x)$
 - a priori knowledge, prior
- Measurement model $P(z|x)$
 - => Likelihood of x given z : $L(x;z) \sim P(z|x)$
- Knowledge **after** measurement: $P(x|z)$
 - a posteriori knowledge, posterior

January 2002

Copyright (c) 2002 Frank Dellaert

44

Yet another way:

- We upgrade our prior (knowledge), by multiplying with the likelihood function, to the posterior (knowledge).

January 2002

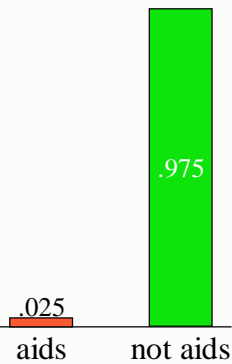
Copyright (c) 2002 Frank Dellaert

45

Applied to 'aids'

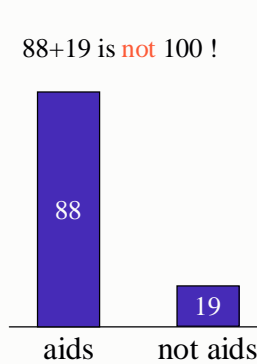
L(aids;pos) ~ P(pos aids)				
aids	test	P(aids, test)	P(aids)	L(aids;pos)
yes	pos	22	25	88%
no	pos	185	975	19%

Prior P(aids)



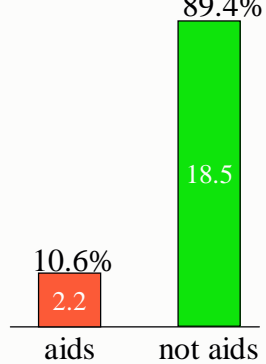
January 2002

Likelihood L(aids;+)



Copyright (c) 2002 Frank Dellaert

Posterior P(aids|+)



46

Outline

Knowledge as Probability

Conditional Probabilities

Probability Densities

Likelihood & Bayes Law

Bayes Classifier

Naïve Bayes Classifier

Binary Classification

1. Pick a prior, e.g. $P(M)=5/6$, $P(F)=1/6$
2. Estimate Measurement Model
 - $P(h|M)=N(h;175,10)$
 - $P(h|F)=N(h;166,10)$
3. Measure person, e.g. height=170
4. Bayes Law:
 - $P(M|170) \sim L(M;170) P(M)$
 - $P(F|170) \sim L(F;170) P(F)$

Choosing a Likelihood Function

- $P(h|M) = N(h; 175, 10)$
= $k \exp(-0.5(h-175)^2/10^2)$
– where $k = (2\pi 10^2)^{-0.5}$
- Likelihood $L(M;h) \sim P(h|M)$
- For example:
– $L(M;h) = \exp(-0.5(h-175)^2/10^2)$
- $P(h|M)$ is Probability, $L(M;h)$ is not !

January 2002

Copyright (c) 2002 Frank Dellaert

49

Bayes Law in Numbers

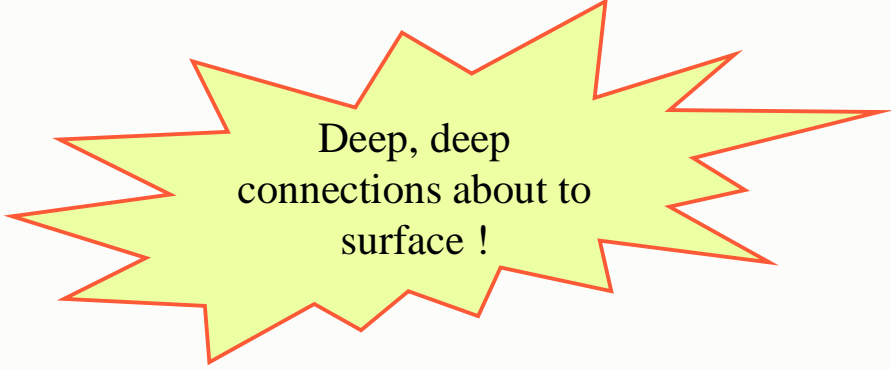
- Bayes Law:
 - $P(M|170) \sim L(M;170) P(M)$
 - $P(F|170) \sim L(F;170) P(F)$
- Plug in Likelihood and Prior:
 - $L(M;170) = \exp(-0.5(170-175)^2/10^2)$
 - MATLAB: `exp(-0.5*(17-17.5)^2)` = 0.8825
 - $P(M|170) \sim 0.8826 * (5/6) = 0.7355$
 - $P(F|170) \sim 0.9231 * (1/6) = 0.1538$

January 2002

Copyright (c) 2002 Frank Dellaert

50

A Mysterious Function



Deep, deep
connections about to
surface !

January 2002

Copyright (c) 2002 Frank Dellaert

51

Posterior as a function of h

- $P(\text{gender}|\text{measured height})$
 - a function of gender !!!
 - but: what happens if we vary h ?
- Posterior = probability
 - Calculate both $\sim P(M|h)$, $\sim P(F|h)$
 - Normalize
 - We have $P(M|h)$!

January 2002

Copyright (c) 2002 Frank Dellaert

52

The form of $P(M|h)$

- $P(M|h) \sim e^{-0.5(h-175)^2/100} P(M)$
- $P(F|h) \sim e^{-0.5(h-166)^2/100} P(F)$
- $P(M|h)$???
- $P(M|h) = \frac{1/[1 + e^{-0.5(h-166)^2/100} P(F)/e^{-0.5(h-175)^2/100} P(M)]}{1/[1 + e^{-s(h-t)}]}$
- Does anybody recognize this ?

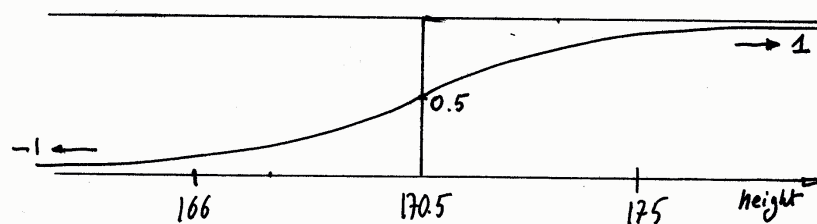
January 2002

Copyright (c) 2002 Frank Dellaert

53

It's the **sigmoid** function

$$\varphi(x;s,t) = 1/[1 + e^{-s(h-t)}]$$



Function central to artificial neural networks !
Basically, ANNs try to learn $P(\text{out}|\text{in})$ **directly**.

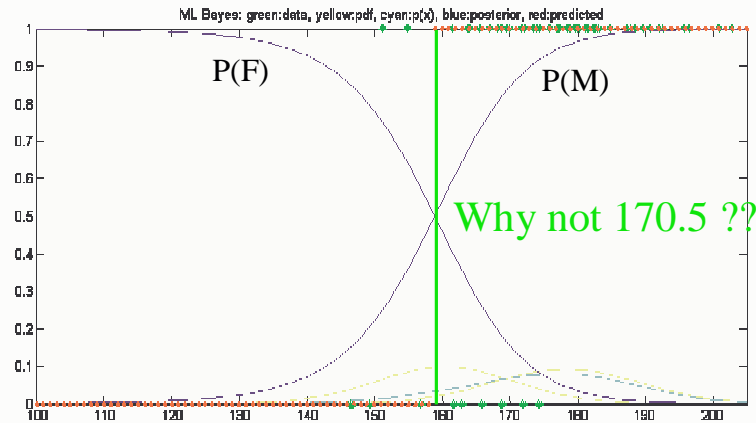
Quiz: is this a probability density function ???

January 2002

Copyright (c) 2002 Frank Dellaert

54

MATLAB example: mlfig04

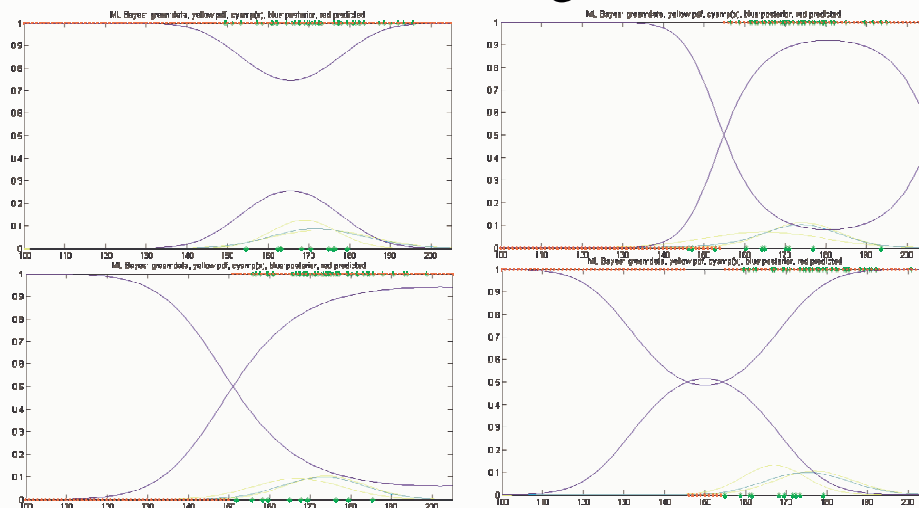


January 2002

Copyright (c) 2002 Frank Dellaert

55

Other training sets:



WHY are they different ? Classification ?

January 2002

Copyright (c) 2002 Frank Dellaert

56

Outline

Knowledge as Probability

Conditional Probabilities

Probability Densities

Likelihood & Bayes Law

Bayes Classifier

Naïve Bayes Classifier

January 2002

Copyright (c) 2002 Frank Dellaert

57

Estimating $P(h|\text{gender})$

- Maximum Likelihood estimate:
 - μ = sample mean
 - σ = (unbiased) sample standard deviation
- Need lots of data when input is high-dimensional !
 - $-\log P(z|x) \sim 0.5 (z - \mu(x))^T \Sigma (z - \mu(x))$
 - μ = sample mean
 - Σ = (unbiased) sample covariance matrix

January 2002

Copyright (c) 2002 Frank Dellaert

58

Naïve Bayes

- Approximate by

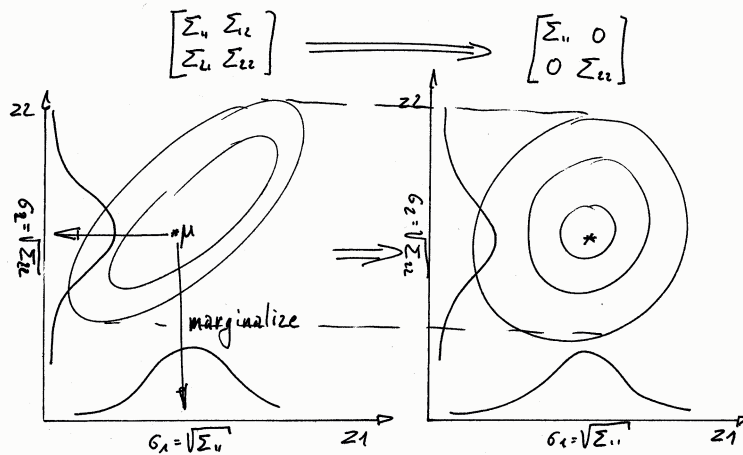
$$P(z_1, z_2, z_3, z_4 | x) \approx P(z_1 | x) P(z_2 | x) P(z_3 | x) P(z_4 | x)$$
- Estimate $N_1(z_1; \mu_1, \sigma_1) \approx P(z_1 | x)$
 - $\mu_1 =$ sample mean z_1
 - $\sigma_1 =$ (unbiased) sample standard deviation z_1
- Ignore all correlations

January 2002

Copyright (c) 2002 Frank Dellaert

59

Naïve Graphics



January 2002

Copyright (c) 2002 Frank Dellaert

60

A Naïve Tennis-player

Mitchell's book p. 179

Estimating Individual Attribute Models:

$$P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{yes}) = 3/9 = .33$$

$$P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{no}) = 3/5 = .60$$

Applying Bayes Law:

$$P(\text{yes}) P(\text{sunny}|\text{yes}) P(\text{cool}|\text{yes}) P(\text{high}|\text{yes}) P(\text{strong}|\text{yes}) = .0053$$

$$P(\text{no}) P(\text{sunny}|\text{no}) P(\text{cool}|\text{no}) P(\text{high}|\text{no}) P(\text{strong}|\text{no}) = .0206$$

Priors

Naïve Likelihood Estimate

January 2002

Copyright (c) 2002 Frank Dellaert

61

The End

January 2002

Copyright (c) 2002 Frank Dellaert

62